Available online at www.sciencedirect.com

ScienceDirect

Applied Mathematics Letters 19 (2006) 1300–1307

Applied Mathematics Letters

www.elsevier.com/locate/aml

Rule learning: Ordinal prediction based on rough sets and soft-computing

P. Pattaraintakorn[a,*], N. Cercone[b], K. Naruedomkul[a]

[a] *Department of Mathematics, Faculty of Science, Mahidol University, Thailand*
[b] *Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada*

Received 18 July 2005; accepted 8 August 2005

Abstract

This work promotes a novel point of view in rough set applications: rough sets rule learning for ordinal prediction is based on rough graphical representation of the rules. Our approach tackles two barriers of rule learning. Unlike in typical rule learning, we construct ordinal prediction with a mathematical approach, rough sets, rather than purely rule quality measures. This construction results in few but significant rules. Moreover, the rules are given in terms of ordinal predictions rather than as unique values. This study also focuses on advancing rough sets theory in favor of soft-computing. Both theoretical and a designed architecture are presented. The features of our proposed approach are illustrated using an experiment in survival analysis. A case study has been performed on melanoma data. The results demonstrate that this innovative system provides an improvement of rule learning both in computing performance for finding the rules and the usefulness of the derived rules.

Keywords: Rough sets; Flow graphs; Rule learning; Soft-computing

Pawlak [1] introduced rough sets theory in the early 1980's. This theory has become recognized and widely researched approaches were conducted [2–5]. Rough sets theory provides mathematicians with the ability to handle uncertainty with approximation. A theoretical point of view of rough sets and related works are provided followed by the central idea behind our proposed designed architecture.

* Corresponding author.
E-mail address: puntip@cs.dal.ca (P. Pattaraintakorn).

0893-9659/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aml.2005.08.004

Applications in a case study of survival analysis and experimental results are presented. We end with concluding remarks and future work.

1. Rough sets theory

In [2–4] the authors proposed exploiting rough sets with relational algebra, decision networks and conflict analysis. Let us explain these ideas and provide a foundation for rough sets considered in these combinations. Assume that there is a finite set $U \neq \emptyset$ called the *universe*. A *decision table* is denoted by $S = (U, C, D)$, where C is the set of *condition attributes* and D is a *target function*. Now rather than consider U, we consider a finite set $\xi \neq \emptyset$ that we define as the *universe of decision rules*. In order to express the relations between C and D with logical formulae, let we consider the logical expressions for a decision rule; $\Phi \rightarrow \Psi$ or "*If* Φ *then* Ψ", where Φ and Ψ are *logical formulae*. Moreover, Φ and Ψ are referred to as *antecedent* and *consequent*, respectively. The *decision table of rules* is defined by $\mathcal{S} = (\xi, \mathcal{F})$, where $\mathcal{F}$ is a set of logical formulae and $\mathcal{R} \subseteq \mathcal{F} \times \mathcal{F}$ is a binary relation, called *consequence relation*.

2. Ordinal prediction

In general, a decision table S is considered to consist of rows labelled by *objects* and columns labelled by *attributes*. The entries in the table are described by *attribute values*. This characterization of attributes uses the notion of attribute values rather than the *ordinal* (criterion), whereas the attribute itself frequently consists of a property: ordinal by nature. For example, any attribute value of the attribute *age* is a continuous value and is implicitly ordered. The consideration of the attribute *age* by values leads us to missing some contexts, e.g., order, scale, hierarchy, increasing or decreasing preference, that are very meaningful in various applications. Especially in medical applications, some attribute values are easier to interpret and obtain knowledge from if they are represented in an ordinal format e.g., the *blood pressure* attribute. We will redefine the ordinal prediction for each decision rule from "*If* C *is* c_1 *then* D *is* d_1" to "*If* C *is in the range between* c_1 *and* c_2 *then* D *is in the range between* d_1 *and* d_2" where c_1, c_2 and d_1, d_2 are values that correspond to attributes C and D, respectively. Given r as the total number of rules, we define an ordinal decision rule R_c, where $1 \leq c \leq r$ as $\Phi_m \rightarrow \Psi_n$, where $\Phi_m : \phi_{m1} \vee \cdots \vee \phi_{mi}$; $\Psi_n : \psi_{n1} \vee \cdots \vee \psi_{nj}$; i, j are the lengths of logical formulae Φ, Ψ, respectively.

3. Rough sets approximations

A family of classifications over ξ is called a *knowledge base*. For $X \subseteq \xi$, if $R \subseteq X \times X$ is an equivalence relation over ξ, then $\xi \setminus R$ means the family of all equivalence classes of R (we will be focus on classification of ξ in our study) referred to as *categories* of R. Let $[x]_R$ denote a category in R containing an element $x \in \xi$. Given a knowledge base $K = (\xi, \mathbf{R})$, $\mathbf{R}$ is a family of equivalence relations over ξ; if $\mathbf{P} \subseteq \mathbf{R}$ and $\mathbf{P} \neq \emptyset$, then there is an equivalence relation $IND(\mathbf{P})$ called the *indiscernibility relation* over $\mathbf{P}$. Having defined $R \in IND(\mathbf{P})$, we have $x = \underline{R}X \leftrightarrow [x]_R \subseteq X$, $x = \overline{R}X \leftrightarrow [x]_R \cap X \neq \emptyset$, called the *R-lower approximation* and *R-upper approximation* of X respectively. Also let $POS_R(X) = \underline{R}X$ denote the *R-positive region* of X, $NEG_R(X) = U - \overline{R}X$ denote the *R-negative region* of X and $BN_R(X) = \overline{R}X - \underline{R}X$ denote the *R-borderline region* of X. We will denote them as *POS*, *NEG* and *BN*, respectively.

4. Rough sets rule categories reduction

In [1] Pawlak proposed to use rough sets in category reduction. In this work we show that this idea can also be expressed differently using rule category reduction for eliminating superfluous rule categories. Given any decision rules denoted by $R_1, \ldots, R_r$, where $R_c \in \xi, 1 \leq c \leq r$ for each consequent Ψ, the category $\xi \backslash R_c$ is *dispensable* in $\cup(\xi \backslash R)$ if $\cup(\xi \backslash R - \xi \backslash R_c) = \cup(\xi \backslash R)$; otherwise the category $\xi \backslash R_c$ is *indispensable* in $\cup(\xi \backslash R)$.

5. Rough sets rule quality measures

In this section, we redefine the traditional card(A) for an ordinal prediction rule, $\text{card}_N(\Phi)$, as the *normalized cardinality* of set Φ. This value means the cardinality of the set in which all elements satisfy Φ in ξ. The meanings of Ψ and $\Phi \wedge \Psi$ are defined in the same manner. Let us define $\text{card}_N(\Phi)$ where i (length of the logical formula $\Phi > 1$) as follows:

$$\text{card}_N(\Phi_1) = \frac{\text{card}(\phi_{11}\vee, \ldots, \phi_{1i})}{i}. \tag{1}$$

Let the goodness of decision rules be measured by the following. The *normalized support* of rule $\Phi \rightarrow \Psi$ is defined as

$$\sup_N(\Phi, \Psi) = \text{card}_N(\Phi \wedge \Psi). \tag{2}$$

Note that we will consider only the rules for which $\sup(\Phi, \Psi) \neq 0$. Moreover, let

$$\text{str}_N(\Phi, \Psi) = \frac{\sup_N(\Phi, \Psi)}{\text{card}_N(\xi)} \tag{3}$$

refer to the *normalized strength* of R. Consequently we have the *normalized certainty* and *normalized coverage* of R as follows:

$$\text{cer}_N(\Phi, \Psi) = \frac{\text{str}_N(\Phi, \Psi)}{\text{card}_N(\Phi)}, \tag{4}$$

$$\text{cov}_N(\Phi, \Psi) = \frac{\text{str}_N(\Phi, \Psi)}{\text{card}_N(\Psi)}. \tag{5}$$

We focus on $\text{card}_N(\Phi) \neq 0$ and $\text{card}_N(\Psi) \neq 0$. In what follows we will use sup, str, cer, cov to denote the values in Eqs. (2)–(5). Moreover, if rule R has cer $= 1$, then R will be called a *certain decision rule* in $\mathcal{S}$; otherwise R will be called an *uncertain decision rule* in $\mathcal{S}$.

6. Rough graphical representation of the rules

Most of the time, the numbers of rules that are constructed from S are very large. For example, more than 1000 rules are constructed for geriatric data in [5]. This number of rules proves problematic since we must select or formulate new rules from the previously large set of rules. Moreover, the results reveal that almost all of the rules are uncertain rules. We will present an idea that can be used as a new approach for data analysis and knowledge representation, by introducing a high level representation of rules. The flow graph representations considered in this section were first proposed by [4]. The example provided in this study is the rough sets application with flow graphs for voting conflict analysis. Nonetheless, the example of a flow graph presented in [4] considered the flow graph with $i = 1$ in Eq. (1).

We will generalize this idea and also complement it with the notion of ordinal prediction. We will redefine the graphical representation of the rules with ordinal prediction while considering the measures in Eqs. (2)–(5).

Let us assume that for every decision table $\mathcal{S}$, there is a *flow graph*, i.e., a directed, acyclic, finite graph associated with $\mathcal{S}$. Given any sequence of antecedents $\Phi_1, \Phi_2, \ldots, \Phi_p$, where $\Phi_k \in \mathcal{F}$, for every $1 \leq k \leq p-1$, $(\Phi_k, \Phi_{k+1}) \in \mathcal{R}$, consider each decision rule as the *path*, $[\Phi_1 \ldots \Phi_p]$, from Φ_1 to Φ_p. We define

$$\mathrm{cer}_N[\Phi_1 \ldots \Phi_p] = \prod_{k=1}^{p-1} \mathrm{cer}_N[\Phi_k, \Phi_{k+1}], \tag{6}$$

$$\mathrm{cov}_N[\Phi_1 \ldots \Phi_p] = \prod_{k=1}^{p-1} \mathrm{cov}_N[\Phi_k, \Phi_{k+1}], \tag{7}$$

$$\mathrm{str}_N[\Phi_1 \ldots \Phi_p] = \frac{\mathrm{card}_N(\Phi_1)\mathrm{cer}_N[\Phi_1 \ldots \Phi_p]}{\mathrm{card}_N(\xi)} \tag{8}$$

$$= \frac{\mathrm{card}_N(\Phi_p)\mathrm{cov}_N[\Phi_1 \ldots \Phi_p]}{\mathrm{card}_N(\xi)}. \tag{9}$$

These paths can form the ordinal prediction rule R_c (from Section 2) as: $\Phi \rightarrow \Psi$ by the *connection*, $\langle \Phi, \Psi \rangle$, from Φ to Ψ. Consequently we define

$$\mathrm{cer}_N\langle \Phi, \Psi \rangle = \sum_{[\Phi \ldots \Psi] \in \langle \Phi, \Psi \rangle} \mathrm{cer}_N[\Phi \ldots \Psi], \tag{10}$$

$$\mathrm{cov}_N\langle \Phi, \Psi \rangle = \sum_{[\Phi \ldots \Psi] \in \langle \Phi, \Psi \rangle} \mathrm{cov}_N[\Phi \ldots \Psi], \tag{11}$$

$$\mathrm{str}_N\langle \Phi, \Psi \rangle = \sum_{[\Phi \ldots \Psi] \in \langle \Phi, \Psi \rangle} \mathrm{str}_N[\Phi \ldots \Psi] \tag{12}$$

$$= \frac{\mathrm{card}_N(\Phi)\mathrm{cer}_N\langle \Phi, \Psi \rangle}{\mathrm{card}_N(\xi)}, \tag{13}$$

$$= \frac{\mathrm{card}_N(\Psi)\mathrm{cov}_N\langle \Phi, \Psi \rangle}{\mathrm{card}_N(\xi)}. \tag{14}$$

7. An example

We are going to provide a second contribution along the lines proposed in [2]. Our system was applied to the melanoma data set (more information appears in [6]). This data is described by seven condition attributes: {*age, sex, ini2, ini3a, ini3b, ini4a, trt*} and a target function: {*nstime*}. The target function is the time to the return to drug use. In our previous study, melanoma data has been analyzed for attribute mining and dimensional reduction [5]. Data cleaning steps are performed to obtain consistent data, then the data were discretized using equal density. Kaplan–Meier survival curves [7] were generated. The effects of all risk factors on the survival curves are compared with log-rank [8], Brewslow [9] and Tarone–Ware tests [10]. These analyses confirm that the risk factors are extracted. Note that from our previous study [2], on rough sets in soft-computing analysis, the *core attributes* = {*age, sex, trt*}. Subsequently, the risk factor that impacts survival time of patients significantly will be considered as a

Table 1
Derived decision rules for melanoma data from HYRIS

Antecedent	Consequent	Cer	Cov	Str
R_1:($age = 6$) and ($sex = 0$) and ($ini3b = 1$)	$nstime = 1$	0.08	0.50	0.05
R_2:($age \leq 3$) and ($ini2 = 1$) and ($trt = 0$)	$nstime = 1$	0.05	0.50	0.05
R_3:($age > 3$) and ($sex = 0$) and ($ini2 = 1$) and ($trt = 0$)	$nstime = 2$	0.06	0.50	0.10
R_4:($age = 4$) and ($sex = 1$) and ($ini3a = 0$) and ($ini4a = 0$) and (trt =0)	$nstime = 2$	0.07	0.25	0.05
R_5:($age = 3$) and ($ini3b = 1$)	$nstime = 2$	0.22	0.25	0.05
R_6:($sex = 1$) and ($ini4a = 1$)	$nstime = 2$	0.17	0.25	0.05
R_7:($age > 3$) and ($sex = 0$) and ($ini2 = 0$) and ($ini3a = 0$) and ($trt = 0$)	$nstime = 3$	0.10	0.25	0.15
R_8:($age = 6$) and ($sex = 1$) and ($ini3b = 1$)	$nstime = 3$	0.09	0.50	0.05
R_9:($age > 4$) and ($sex = 1$) and ($ini2 = 1$)	$nstime = 4$	0.05	0.33	0.05
R_{10}:($age = 1$) and ($trt = 0$)	$nstime = 4$	0.10	0.33	0.05
R_{11}:($age > 4$) and ($ini2 = 1$) and ($trt = 1$)	$nstime = 4$	0.06	0.33	0.05
R_{12}:($age = 1$) and ($trt = 1$)	$nstime = 5$	0.18	0.50	0.05
R_{13}:($3 \leq age \leq 4$) and ($ini3a = 1$)	$nstime = 5$	0.12	0.33	0.05
R_{14}:($age \leq 3$) and ($sex = 1$)	$nstime = 6$	0.13	0.67	0.10
R_{15}:($age \leq 3$) and ($ini3b = 0$) and ($trt = 1$)	$nstime = 6$	0.06	0.33	0.05
R_{16}:($age > 4$) and ($ini3a = 1$)	$nstime = 6$	0.08	0.33	0.05
Average length of the rules = 3.06	Average	0.20	0.38	0.06

probe attribute (defined in [2]). Finally, we consider all test measures and the core attribute together as the *probe attribute* = {*ini3b*}. The *probe reducts* = {*age, sex, trt, ini2, ini3a*}. The ordinal decision rules for predicting survival tendency for melanoma data are generated with ELEM2 [11]. Table 1 shows an example of 16 derived diagnosis rules from melanoma data with 100% accuracy.

These form a very small number of example rules. Nonetheless, the structures of the rules are difficult to obtain significant knowledge from or even interpret. Furthermore, all rules are uncertain rules because cer $\neq 1$. We perform rule category reduction following Section 4, resulting in discarding R_2 and R_6. Next, we find the highest normalized strength $\mathrm{str}_N(\Phi, \Psi)$ for each consequent and result as the attribute *age*. We then use *age* and *nstime* as a knowledge base and find *POS*, *NEG* and *BN* to represent the rules R_1, R_3 to R_5, R_7 to R_{16} as shown in Table 2. These *POS*, *NEG* and *BN* are used to select the informative rules together with normalized certainty of each connection. The symbols in Table 3, '+', '−' and 'o', denote that the rule R can be used to predict that an example belongs to the positive region, negative region and borderline region with respect to the knowledge base *age, nstime*, respectively. The candidate rules that have the most predictive powers are listed: $\{R_1, R_4, R_5, R_8, R_9, R_{10}, R_{11}, R_{12}, R_{15}, R_{16}\}$. Note that R_{15} is included because we want to investigate the probe attribute *ini3a*. The flow graphs of these selected rules are constructed and the relationships among attributes *age*, *sex, ini2*, *ini3b*, *trt* and *nstime* are depicted with the flow graph of Fig. 1. Each node represents logical formula Φ or Ψ, each flow represents a consequence relation and each antecedent connected to the *nstime* logical formula represents a $\langle \Phi, \Psi \rangle$ connection. Only one attribute, *age*, in this figure has the cardinality of the logical formula greater than 1. The measures in Eqs. (1)–(14) are calculated to analyse this flow graph.

According to Fig. 1, we can construct the ordinal prediction. The attribute *age* consists of three groups: {1, 2}, {3, 4}, {5, 6} and *nstime* consists of four groups: {1}, {2, 3}, {4, 5}, {6}. Hence the ordinal prediction rules with attribute values can be derived as shown in Table 4.

Table 2
Rough representations of the rules with respect to the knowledge base *age, nstime*

nstime : *POS*	*NEG*	*BN*
1:$R_1, R_5, R_8, R_{10}, R_{12}, R_{14}, R_{15}$	$R_3, R_7, R_9, R_{11}, R_{13}, R_{16}$	R_4
2:$R_1, R_3, R_4, R_5, R_7, R_8, R_9, R_{11}, R_{13}, R_{16}$	R_{14}, R_{15}	R_{10}, R_{12}
3:$R_1, R_3, R_4, R_7, R_8, R_9, R_{11}, R_{16}$	R_{13}	$R_5, R_{10}, R_{12}, R_{14}, R_{15}$
4:$R_1, R_8, R_9, R_{10}, R_{11}, R_{12}, R_{16}$	R_3, R_7, R_{14}, R_{15}	R_4, R_5, R_{13}
5:$R_4, R_5, R_{10}, R_{12}, R_{13}$	R_3, R_7, R_{14}, R_{15}	$R_1, R_8, R_9, R_{11}, R_{16}$
6:$R_1, R_5, R_8, R_9, R_{10}, R_{11}, R_{12}, R_{14}, R_{15}, R_{16}$	R_3, R_7, R_{13}	R_4

Table 3
Candidate rules selected from the rough representation

Age	R_1	R_2	R_3	R_4	R_5	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}
1	+	+	0	−	+	0	+	0	+	0	+	0	+	+	0
2	+	0	+	+	+	+	+	+	−	+	−	+	0	0	+
3	+	−	+	+	−	+	+	+	−	+	−	0	−	−	+
4	+	0	0	−	−	0	+	+	+	+	+	−	0	0	+
5	−	0	0	+	+	0	−	−	+	−	+	+	0	0	−
6	+	+	0	−	+	0	+	+	+	+	+	0	+	+	+
Cer	0.08	0.05	0.06	0.07	0.22	0.10	0.09	0.05	0.13	0.06	0.18	0.12	0.13	0.06	0.08

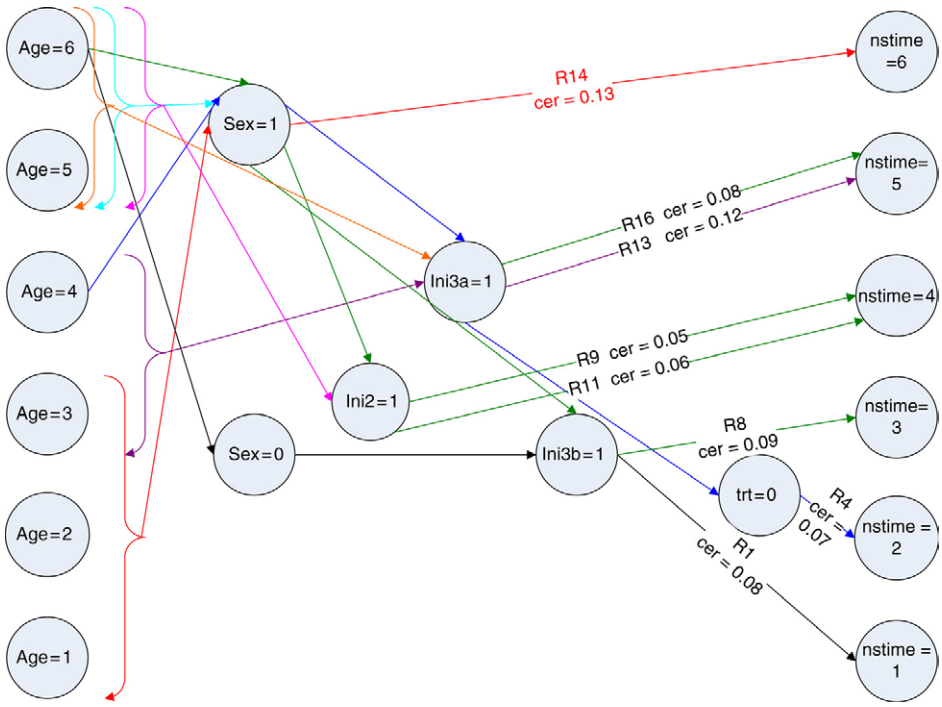

Fig. 1. Flow graph rule representation.

Table 4
Ordinal prediction rules

Number	Rules		Cer	Cov	Str
1.	If $18 \leq age \leq 35$ and $sex = 1$	then $24 < nstime \leq 34$	0.18	0.50	0.10
2.	If $35 < age \leq 49$ and $ini3b = 1$	then $1 < nstime \leq 7.8$	0.27	0.44	0.20
3.	If $49 < age \leq 91$ and $ini3b = 1$	then $1 < nstime \leq 14$	0.06	0.09	0.05
4.	If $49 < age \leq 91$ and $ini2 = 1$	then $14 < nstime \leq 34$	0.27	0.44	0.20
Average length of the rules = 2.00		Average	0.20	0.37	0.17

The resulting rules present us with a greatly reduced number of rules: from 16 unique prediction rules to 4 ordinal prediction rules. The average length of the rules also reduces from 3.06 to 2.00. While keeping the same values for cer and cov, when compared with the results from Table 2, the average str for rules increases significantly from 0.06 to 0.17.

8. Concluding remarks and future work

Our approach illustrates the formulation of more meaningful rules using the notion of ordinal prediction. The results are the rules that are constructed from the interval antecedents and are able to predict intervals rather than unique values of the target function. Furthermore, each decision rule will be represented by its rough representation. Our innovative approach proved to be an improvement for rule learning both in computing performance and the usefulness of the rules derived from a case study on melanoma data. Our future works will provide a systematic approach for rule induction from the flow graph.

Acknowledgements

This research was supported by NSERC, UDP, Mahidol University and KMITL. Thanks are also due to Arnold Mitnitski, S. Jiampojamarn, G. Zaverucha. The first author was partially supported by UDP, Thailand. The second author was partially supported by NSERC, Canada.

References

[1] Z. Pawlak, Rough sets, in: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.
[2] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Hybrid intelligent systems: Selecting attributes for soft-computing analysis, in: Proc. of 29th Proc. Annual International COMPSAC, 2005 (in press).
[3] Z. Pawlak, Decision networks, Proc. of the Forth International Conference, RSCTC, in: S. Tsumoto, R. Slowinski, J. Komorowski, J. Grzymala-Busse (Eds.), in: LNAI, vol. 3066, 2004, pp. 1–7.
[4] Z. Pawlak, Some remarks on conflict analysis, European J. Oper. Res. 166 (2005) 649–654.
[5] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Selecting attributes for soft-computing analysis in hybrid intelligent systems, in: Proc. of the Tenth International Conference RSFDGrC, 2005 (in press).
[6] T.E. Lee, J.W. Wang, Statistical Methods for Survival Data Analysis, 3rd edition, John Wiley & Sons, 2003.
[7] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Amer. Statist. Assoc. 53 (1958) 457–481.
[8] A.V. Peterson Jr., Expressing the Kaplan–Meier estimator as a function of empirical subsurvival functions, J. Amer. Statist. Assoc. 72 (1977) 854–858.

[9] R. Peto, J. Peto, Asymptotically efficient rank invariant procedures, J. R. Stat. Soc. 135 (1972) 185–207.
[10] E.A. Gehan, A generalized Wilcoxon test for comparing arbitrarily singly-censored data, Biometrica 52 (1965) 203–223.
[11] A. An, N. Cercone, ELEM2: A learning system for more accurate classifications, in: Proc. of CSCSI, in: LNCS, vol. 1418, Springer-Verlag, London, 1998, pp. 426–441.